Unraveling the Structure of Knowledge: Consistency in Everyday Networks, Diversity in Scientific

Owen G. W. Saunders, Chico Q. Camargo, and Massimo Stella

Abstract The patterns and dynamics that govern the flow of concepts and the association of information for various scientific domains, are not well understood in the realm of knowledge evolution and organisation. To that end, we look at using concept networks to capture the associations between these concepts as a domain grows. We compare concept networks as they grow for scientific domains, sci-fi literature, common news topics and science news, using Quantum Spectral Jensen-Shannon Divergence, to evaluate how consistent their network structures are at their early stages. We find that everyday concept networks tend to be more consistent with each other, whereas scientific networks are less consistent and we discuss the potential factors influencing the structures of these networks.

Keywords Science of science · Network evolution and growth · Concept networks · Divergence

1 Introduction

It is essential to consider how scholars are exploring, developing and creating concepts and ideas. For example, closely related ideas within a scholar's expertise can facilitate a more focused and stable approach to research, whereas engaging with broader concept exploration brings a higher chance of making major breakthroughs but could encourage riskier careers [2]. These webs of ideas can be modeled as con-

O. G. W. Saunders (✉) · C. Q. Camargo
Faculty of Environment, Science and Economy, University of Exeter, Stocker Rd, Exeter EX4 4PY, UK
e-mail: os318@exeter.ac.uk

C. Q. Camargo
e-mail: f.camargo@exeter.ac.uk

M. Stella
CogNosco Lab, Department of Psychology and Cognitive Science, University of Trento, Rovereto, Trento 38068, Italy
e-mail: massimo.stella@inbox.com

F. Botta et al. (eds.), *Complex Networks XV*, Springer Proceedings in Complexity,
https://doi.org/10.1007/978-3-031-57515-0_10

cept networks, where nodes of the network are ideas or words, and edges between nodes represent the association and co-occurrence of ideas in a domain, being particularly useful in prior studies for tasks such as link prediction [5, 12].

Within the field of Science of Science however, the study of such concept networks aims to reveal underlying patterns and dynamics governing the diffusion and evolution of knowledge, offering insights into how ideas and concepts interconnect, propagate, and influence various scientific domains. The focus of this study is the structural characteristics of the concept networks instead of their actual contents, although previously it has been shown that there is a connection between the contents of concept networks and their structure, and further that the structural features of a network can be connected to the quality and coherence of the contents, i.e. the logical flow and consistency of ideas [4].

This study thus has the goal of comparing the structures of scientific concept networks as they grow over time, intending to provide a clearer picture of how knowledge flows within these scientific domains, and whether they show common structural patterns, or develop major differences in topology during their initial formation. Studies on the differences between networks have already been carried out by many researchers [1, 3, 9], with multiple measures to choose from. However, the analyses for a network are highly dependent on the selected analytical tool, and since we wish to perform pairwise comparisons of networks, the most critical choice is which methodology to use for calculating the distances between them [1].

The distance measures provided by current literature capture distinct properties, and can be roughly separated into structural versus spectral distances. Structural distances show local changes, but are not suited for our comparisons as they fail to capture the overall and specific changes in information between networks e.g. Hamming distance only shows us the amount of change between two networks, since it is a special instance of the broader Graph-Edit distances—see Fig. 1 which shows the tendency towards zero for our sparse concept networks. Spectral distances on the other hand are global measures that use the eigenvalues from either a version of

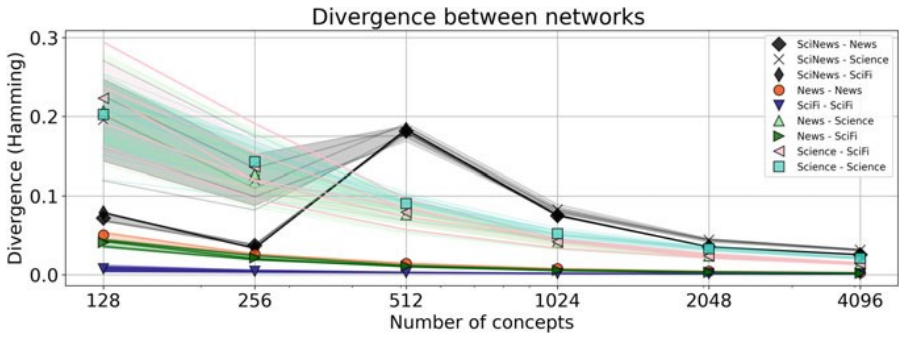


Fig. 1 Hamming divergence for pairwise comparisons between descriptors of concept networks with increasing numbers of concepts. *Note* Networks of different sizes are not compared to each other

the Laplacian L or the adjacency matrix A, and are more suited for this study because the eigenvalues of a network are good descriptors for the topological structure, with the additional benefit that spectral distances do not take into account the identities of nodes [1].

The remainder of this paper is split into three sections as follows: Sect. 2 discusses the methods used, from the setup of our concept networks to the datasets, problem formulation and distance measure used, shortly followed by potential factors influencing the structures of the concept networks. Section 3 presents the results alongside a discussion for the hypotheses previously presented. Section 4 concludes this paper and highlights future work and challenges.

2 Methods

2.1 Data Collection and Network Creation

We perform network analyses for a variety of texts, constructing single-layer concept networks for scientific domains, news categories and science fiction literature, where nodes represent concepts extracted from the abstract, news article summary or lines from the first chapter of a science fiction book, and edges represent co-occurrence of concepts within those sources (e.g. co-occurrence of scientific terms within a single abstract). We take abstracts from arXiv preprints [10], news articles from the HuffPost News Category Dataset [7, 8] and IEEE Dataport [11], and science fiction literature from the SF Nexus Corpus (note: this dataset only exhibits extracted features for copyrighted fiction; i.e., no copyrighted work was made available for consumption) [14]. Science Fiction and news articles provide diverse topics over multiple genres and high regularity/structure respectively, which leads them to be the most suitable for our comparisons. To ensure that an adequate sample is produced from the Science Fiction literature, we sample concepts from books over multiple selected decades from 1920 to 2010.

Concepts are obtained using simple natural language processing (NLP) and entity recognition (ER), with no concepts excluded to avoid removing words based on localised preconceptions, e.g. removing the word force as a non-concept in one domain ("too much force") compared to in the domain of physics ("the force in Newtons is..."). Concepts from abstracts and science fiction are extracted using a Large Language Model (LLM) trained on scientific words, obtained from the SciSpaCy[1] project which provides SpaCy[2] models for processing biomedical, scientific and clinical texts. The accuracy is very high for this model, but it has been trained

[1] https://allenai.github.io/scispacy/.

[2] https://spacy.io/.

Table 1 Overview of basic network properties for the concept networks created

Concept networks[a]	Edges	Components			
		Giant[b]	Giant %	Second largest	All connected
scinews	211297	4009	97.8	2	16
base_news_business	15790	3730	91.1	2	125
base_news_comedy	10808	3165	77.3	2	361
base_news_home_and_living	16490	3785	92.4	2	74
science_cs_CG	106313	4096	100	0	1
science_econ_TH	116410	4096	100	0	1
science_physics_space-ph	124820	4096	100	0	1
science_q-bio_OT[c]	122008	4092	99.9	4	2
science_stat_OT	107779	4096	100	0	1
scifi_1920s	4708	3283	80.2	2	361
scifi_1940s	3892	2675	65.3	2	597
scifi_1950s	3878	2537	61.9	2	653
scifi_1960s	3784	2596	63.4	2	627
scifi_1980s	4024	2703	66.0	2	598
scifi_1990s	3761	2541	62.0	2	665
scifi_2000s	4847	3357	82.0	2	324
scifi_2010s	5021	3281	80.1	2	361

[a] All Concept Networks are 4096 nodes in size
[b] The Giant component is the largest component of a network
[c] The second component here is from part of the last paper added to this concept network, which did not mention any of the other 124 concepts

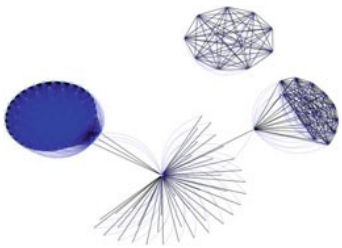
and tested for biomedical corpora so may experience a decrease in accuracy for other domains, however this is negligible enough for us to ignore for the purposes of this study.

To allow for the chronological addition of concepts when we use networks of varying sizes, each concept-concept edge must be assigned a timestamp, obtained from the source's published date and time. We observe networks of up to 4096 nodes, limiting the networks to only the concepts used in the initial formation stages of the scientific domains and their knowledge structures.

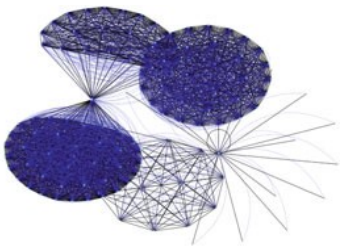
Table 1 shows the most relevant properties—edges and connected components—for the concept networks at their largest size of 4096 nodes. We can immediately see that the science domains tend to form a single large giant component that represents a highly interconnected web as shown by the edge count, although this is comparative to the non-scientific domains, as even with 100,000 edges, this is approximately only 1% of the possible 8.4 million edges in a fully connected network of 4096 nodes.

2.2 Concept Networks

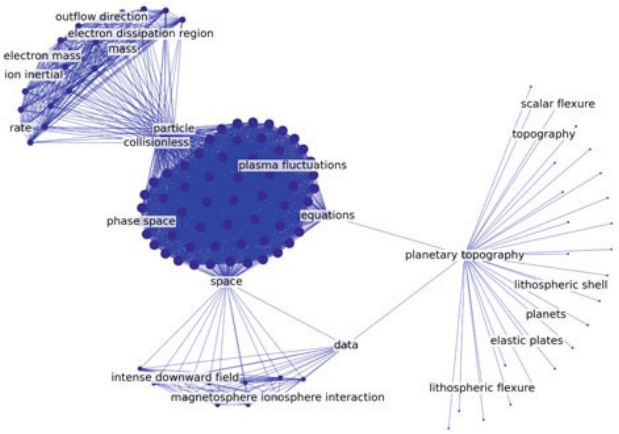
The concept networks produced for science tend to exhibit highly connected clusters that are separated by certain bridging concepts, shown in Fig. 2. These bridging concepts are words such as "background" in the quantitative biology network drawn in Fig. 2a. Although the scientific domains all have very large giant components, this is partially due to these bridging concepts which are reused throughout multiple papers as they are considered general and common scientific words, viz. data, model, result, background, algorithm etc.



(a) science_q-bio_OT

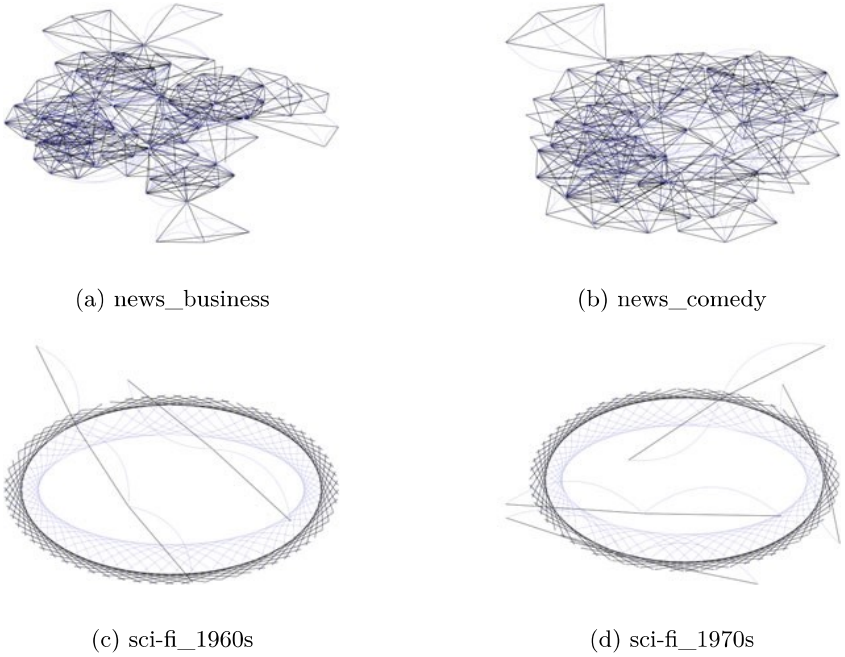


(b) science_stat_OT



(c) science_physics_space-ph

Fig. 2 Three concept networks for selected science domains at sizes of 128 concepts. **Top Left and Right**: Both black and blue edges show co-occurrence, but blue edges are force directed to show bright blue sections in dense interconnected regions. **Bottom**: A partially labelled concept network for example purposes, where node size represents degree. Note that common terms tend to act as bridges between communities, viz. space, equations, particle



(a) news_business (b) news_comedy

(c) sci-fi_1960s (d) sci-fi_1970s

Fig. 3 Four concept networks consisting of two news categories, and two decades of sci-fi literature at sizes of 128 concepts. Both black and blue edges show co-occurrence, but blue edges are force directed to show bright blue sections in dense interconnected regions. Note that the science fiction networks only consist of **disconnected** 1-, 2- and 3-degree nodes at small sizes, due to the filtering for scientific words only

We've chosen to keep these common words to prevent any biases from occurring from manually removing or labelling each concept as scientific or general, however research on language kernels may be able to automatically and systematically define common and general 'throwaway'/definition words in science, provided we have definitions for the necessary words to remove [13].

In contrast, the news and sci-fi domains in Fig. 3 show more consistent structures, with news categories generally having many small communities around different topics that have popped up in articles and sci-fi literature showing many groups of 2–5 nodes, which is due to the usage of the LLM for sci-fi literature. These sci-fi networks are more prone to structural change with larger sizes due to low initial connectivity and an increase in authors after including more books.

The science news concept network in Fig. 4 is an intuitive mid-ground between the topology of the news and science networks. It depicts around 6–10 clear communities which are focused on different topics/domains from the news, but still maintains clear bridging concepts between these topics. It must be noted that some concepts in this network are more related to business than science, e.g. tech companies or policies can be recurring themes, although these still count as important concepts of science news so they have not been removed.

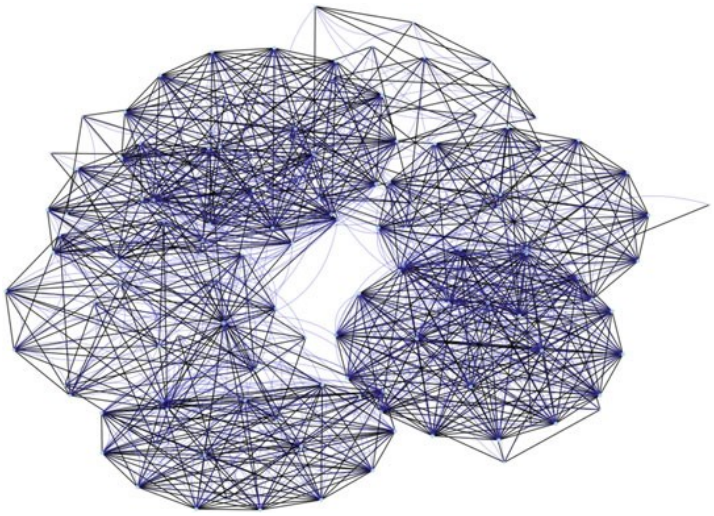


Fig. 4 The science news concept network at a size of 128 concepts. Both black and blue edges show co-occurrence, but blue edges are force directed to show bright blue sections in dense interconnected regions

2.3 Divergence Metric

To compare these networks we employ divergence metrics, using unweighted adjacency matrices with Quantum Spectral Jensen Shannon Divergence (QJD), a normalised and symmetric metric for spectral entropy, the implementation for which we take from the netrd[3] package [6]. This allows for comparisons between networks of different sizes if needed and is the best choice for cost-performance constraints in respect to time complexity and results. Note that these particular mathematical descriptions have been previously demonstrated to effectively capture established graph properties, based on research employing random graph models [3]. Furthermore, QJD is graph invariant, i.e., it uses representations that are universally applicable to all isomorphs of a graph.

2.4 Research Hypotheses

Given our chosen divergence metric and methodology, and the preliminary example concept networks displayed so far, we propose the following hypotheses to test, where $d()$ represents our divergence metric:

[3] https://github.com/netsiphd/netrd.

Proposition 1 *In general, the distance between any two science domains, e.g. biology and physics, should be greater than the distance between any two news domains, and both distances should be greater than that between any two decades of sci-fi literature:*

$$d(science_i, science_j) > d(news_k, news_l) > d(scifi_m, scifi_n) \tag{1}$$

Proposition 2 *In general, the distance between any two news domains should be less than the distance between any two science news domains, and both distances should be less than that between any two domains of science:*

$$d(news_i, news_j) < d(science\_news_k, science\_news_l) < d(science_m, science_n) \tag{2}$$

Proposition 3 *In general, the distance between any two domains of science should be more similar than the distance to other domains:*

$$\begin{aligned} d(science_i, science_j) - d(science_k, science_l) < d(domain_i, domain_j) \\ - d(domain_k, domain_l) \end{aligned} \tag{3}$$

where domains $i, j, k, l \notin science$.

Proposition 4 *As a null test, the distance between any two concept networks of the same domain (not subdomain/category), viz. science, news, sci-fi, should be less than 0.5, i.e. they are considered to have consistent topologies:*

$$d(domain_i, domain_j) < 0.5 \tag{4}$$

Proposition 4 is expected to be false as we can see easily even for the 128-node concept networks, that scientific domains vary greatly.

3 Results and Discussion

We take the null hypothesis that the structures of these concept networks will be consistent, viz. the distance for any pairwise combination is closer to 0 than to 1 as defined in Proposition 4. This null hypothesis is rejected, and we find that there is little consistency in the structure of the networks of ideas and concepts used across different scientific domains, as shown in Fig. 5. In contrast, there is a lot more consistency in structured sources of everyday language such as news articles or literature.

For the rest of our predictions, Proposition 1 can be accepted, as we can see distinctly the difference in consistency both in the figures of concept networks and in Fig. 5, where our Sci-Fi, News and Science domains stay separated by a large margin even when increasing in size dramatically.

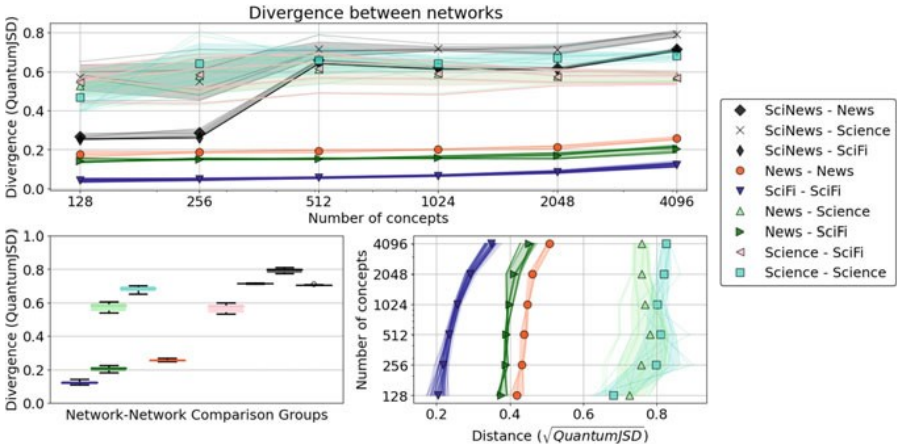


Fig. 5 Quantum Spectral Jensen-Shannon Divergence (QJD) as calculated for pairwise combinations of each network (e.g. biology versus physics). **Top**: All networks. **Right**: Intra-distances between Science, SciFi and News networks and inter-distances for (Science, News) and (News, SciFi) **Left**: Boxplot comparisons for 4096 concept networks only

Table 2 Network properties for news concept networks

Network size	Business news		Science news[a]	
	Edges	Connected components	Edges	Connected components
128	382	13	682	12
256	731	25	1132	13
512	1387	44	**24244**	15
1024	2816	70	**39829**	10
2048	6059	105	**74042**	17
4096	15790	125	**211297**	16

[a]Note the sudden increase in connectivity. This is explained in the results and discussion

Proposition 3 holds for all domains except the Science News, so it is accepted, since Science news is considered as belonging to a scientific domain.

However Proposition 2 showed unexpected results. For 128 and 256 node concept networks the results were as predicted for the distance between science news and other news categories, but between 256 and 512 nodes the structure becomes a lot more inconsistent with the other news categories. With further investigation into the changes in network properties, it is seen in Table 2 that the science news concept network becomes suddenly more interconnected. This is due to a temporal limitation of the Science and Technology News dataset. Since the data collection period was shorter and more compact than the other news datasets, it resulted in a narrow focus of topics. The 512 node concept network had high degree nodes such as “ireland”, “dublin city”, “corporate” and “saturday”, showing that most of the news articles were centred around a specific event or prominent topic.

4 Conclusion and Future Work

These analyses demonstrate that it is possible to perform meaningful large-scale comparisons of how information is structured in different domains, scientific or not. We show that whilst there are certain similarities between the structure of ideas and concepts in scientific domains, a specific layout and ordering of information is not crucial to the development of a newly created scientific field in terms of connecting ideas. Concept networks for everyday domains tend to have consistent structures with each other, whereas concept networks for science tend to be less consistent with each other, but still retain some similar features such as generic and domain specific bridging concepts.

Future work with particular relevance to this paper would look at:

- Whether or not science news of a particular scientific domain is consistent, in terms of concept network structure, with the domain of science it is trying to communicate about.
- Comparisons of graphs using meso-scale dissimilarity measures to capture the differences in communities, instead of the networks as a whole, could provide more meaningful comparisons for the structuring of subtopics and areas of research within a domain.
- Exploration of concepts within these networks and semantic spaces, looking at the actual identities of concepts in these networks and whether or not there may be 'research gaps' which can be identified based on the network structure and contents.
- Validation of some of the suggestions raised in discussion of our results, such as whether or not scientific news develops a web like structure similar to domains of science, if given data spanning a similar time frame.
- Changes in the number of "bridging concepts" in science domains over multiple decades.

Lastly, it is interesting to note that sci-fi domains had 'chains' of concepts, news domains had 'rings' and 'necklaces' and science domains formed interconnected 'webs'. This fits particularly well with research published by Koponen and Pehkonen on coherence and conceptual networks [4], so certain correlations in these base structural relations could be investigated.

References

1. Donnat, C., Holmes, S.: Tracking network dynamics: a survey using graph distances. Ann. Appl. Stat. **12**(2), 971–1012 (2018)
2. Fortunato, S., Bergstrom, C.T., Börner, K., Evans, J.A., Helbing, D., Milojević, S., Petersen, A.M., Radicchi, F., Sinatra, R., Uzzi, B., et al.: Science of science. Science **359**(6379), eaao0185 (2018)

3. Hartle, H., Klein, B., McCabe, S., Daniels, A., St-Onge, G., Murphy, C., Hébert-Dufresne, L.: Network comparison and the within-ensemble graph distance. Proc. R. Soc. A **476**(2243), 20190744 (2020)
4. Koponen, I.T., Pehkonen, M.: Coherent knowledge structures of physics represented as concept networks in teacher education. Sci. Educ. **19**, 259–282 (2010)
5. Krenn, M., Buffoni, L., Coutinho, B., Eppel, S., Foster, J., Gritsevskiy, A., Lee, H., Lu, Y., Moutinho, J., Sanjabi, N., Sonthalia, R., Tran, N., Valente, F., Xie, Y., Yu, R., Kopp, M.: Predicting the future of AI with AI: high-quality link prediction in an exponentially growing knowledge network. Nat. Mach. Intell. (2022)
6. McCabe, S., Torres, L., LaRock, T., Haque, S.A., Yang, C.H., Hartle, H., Klein, B.: netrd: a library for network reconstruction and graph distances (2020). arXiv:2010.16019
7. Misra, R.: News category dataset (2022). arXiv:2209.11429
8. Misra, R., Grover, J.: Sculpting Data for ML: The First Act of Machine Learning (2021)
9. Scott, C., Mjolsness, E.: Graph diffusion distance: properties and efficient computation. PLoS ONE **16**(4), e0249624 (2021)
10. arXiv.org Submitters: arxiv dataset (2023). https://www.kaggle.com/dsv/6697593
11. Thakur, R.: Science and tech news dataset (2021). https://doi.org/10.21227/wtzb-0w91
12. Tran, N.M., Xie, Y.: Improving random walk rankings with feature selection and imputation science4cast competition, team hash brown. In: 2021 IEEE International Conference on Big Data (Big Data), pp. 5824–5827. IEEE (2021)
13. Vincent-Lamarre, P., Massé, A.B., Lopes, M., Lord, M., Marcotte, O., Harnad, S.: The latent structure of dictionaries. Top. Cogn. Sci. **8**(3), 625–659 (2016)
14. Wermer-Colan, A.: Sf corpus hugging face (2023). https://huggingface.co/datasets/SF-Corpus/